ALY6000

Probability Theory and Introductory Statistics

Module 6 – R Practice 6

Name: Duc Viet Mai

Student ID: 002661113

Date: December 15, 2022

I. INTRODUCTION

Using the same dataset testing the effect of stroke-preventing drugs - Gabapentin, I have conducted several regression analyses using dummy variables on two subsets: one with the use of Gabapentin and one with a placebo. In the following section, I will further explain my dependent and independent variables and what I expect to gain from the results.

II. DATA ANALYSIS

- **Part 1**

In the first part, I created a dummy variable called "intervention" that has the value of 1 if the treatment is "Gabapentin" and 0 if the treatment is "Placebo." Then, I reran the regression model with the dependent variable "Outcome," the independent variable "White blood cells," and a dummy variable "intervention."

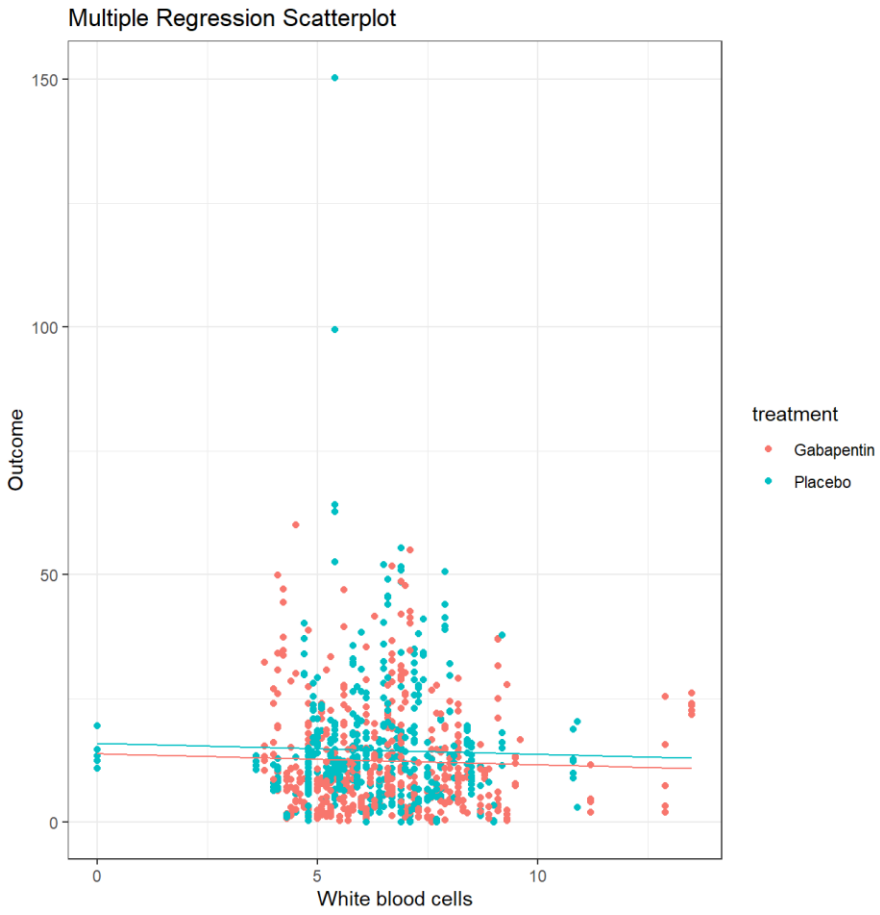
```
Call:
lm(formula = outcome ~ wbc + intervention, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-14.539  -8.052  -3.126   3.981 135.607

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.8874     1.5951   9.960   <2e-16 ***
wbc           -0.2211     0.2358  -0.938   0.3486
intervention  -2.1075     0.7838  -2.689   0.0073 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.12 on 958 degrees of freedom
Multiple R-squared:  0.008804,  Adjusted R-squared:  0.006734
F-statistic: 4.254 on 2 and 958 DF,  p-value: 0.01447
```

Looking into the summary of the model, you can notice that there will be two equations to calculate the predicted value y, which I will present below. The main difference between the two equations is one will take into account the treatment with Gabapentin drugs, while the other will not due to placebo treatment. Specifically, you can observe this vital difference in the coefficient value of the dummy variable, which is equal (-2.1075). Henceforth, the predicted value y in equation 1 will be higher in equation 2. In detail, because equation 2 has the dummy variable's indicator of 1 (not 0 because this is not the placebo group), multiplied by (-2.1075), which equals the predicted value in equation 1 minus (1*-2.1075), resulting in the lower value y in equation 2. The interpretation would be one unit change in the treatment results in a decrease of 2 units in the outcome.

$$Equation\ 1: \hat{y} = 15.8874 - 0.2211 * wbc - 2.1075 * 0$$

$$Equation\ 2: \hat{y} = 15.8874 - 0.2211 * wbc - 2.1075 * 1$$

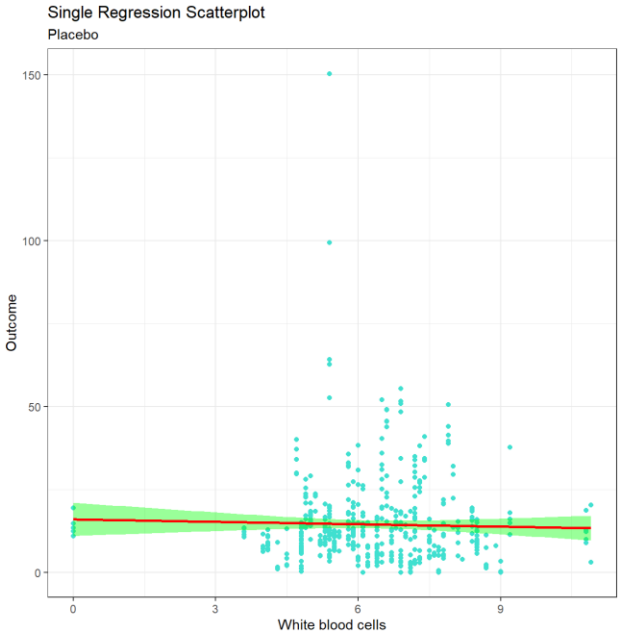
Furthermore, you can see this difference in the variation of the data points is a difference on the scatterplot, with the red linear regression (Gabapentin) lower than the green one (Placebo).

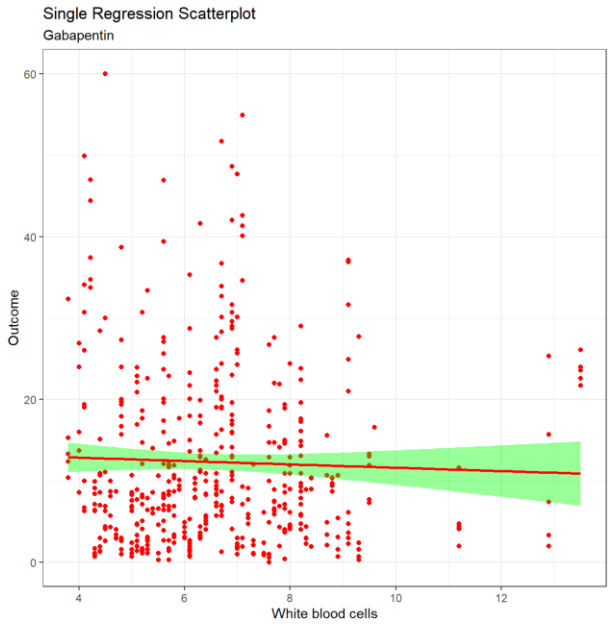


Overall, in this multiple regression analysis, there are two subsets: one with Gabapentin drugs (dummy variable "intervention" = 1) and the other with a placebo (dummy variable "intervention" = 0). There are two lines representing each linear regression in each group. And by visualizing the regression model, we acknowledge that the impact of the categorical variable "treatment" does shift our predicted value y. Adding the dummy variable will increase the validity of our prediction, and we might be aware of any confounding factors. In the future, we should exploit this method to reassure the precision of our predicted results.

- **Part 2**

The separated regression line in the scatterplot for the two above subsets is the same within the multiple regression analysis. However, by looking at them individually, we can see more clearly the impact of sub-setting our data sets based on the dummy variable. Each subset will have its regression model with a different predicted value.





Specifically, the first subset with Gabapentin has lower white blood cells, in general, results in a lower predicted outcome (maximum value is around 60). In contrast, the second subset with placebo has higher white blood cells, showing a higher outcome value, with a maximum of more than 100.

REFERENCES

[1] Bluman, A. G. (2018). *Elementary statistics: A step by step approach.* McGraw-Hill Education. 10th ed, 1- 368. ISBN: 978-1-259-75533-0.

[2] Kumar, A. (April 20, 2022). *Dummy variables in regression models: Python, R.* Data Analytics. Retrieve from: https://vitalflux.com/dummy-variables-in-regression-models-python-r/

[3] Moon, K.W. (October 6, 2020). *ggPredict() - Visualize multiple regression model.* Retrieve from: https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html

[4] Statistical tools for high-throughput data analysis (STHDA). *Ggplot2title: main, axis and legend titles.* Data Visualization. Retrieve from: http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles

[5] Zach. (February 2, 2021). *How to create dummy variable in R.* Statology. Retrieve from: https://www.statology.org/dummy-variables-in-r/